**Universidade de São Paulo**

**Faculdade de Saúde Pública**

**Departamento de Epidemiologia**

**Disciplina HEP165 - 2023**

STATA

Professora: Denise Pimentel Bergamaschi

2023

**Orientações para o uso do material:**

Esse material é um roteiro para as aulas de disciplina oferecida na Faculdade de Saúde Pública da USP e aborda a aplicação do pacote de análise de dados estatísticos Stata (Stata, Coorp, Versão ).

Propõe-se que o material seja aberto como documento texto para que seja modificado com as anotações e comentários que o aluno julgar necessários durante as aulas, assim o material ficará mais completo e ajustado ás necessidades de aprendizado de cada um.

O material contém, no final, uma sessão com o gabarito das questões.

**Para tornar o uso do material mais fácil, notar que:**

O nome das variáveis estará em *itálico*

Os comandos serão apresentados em ***negrito e itálico***

Os nomes de arquivos estarão entre os sinais <.......>

É possível usar o menu do Stata e este será indicado da seguinte forma: File > Open...

O que é o Stata?

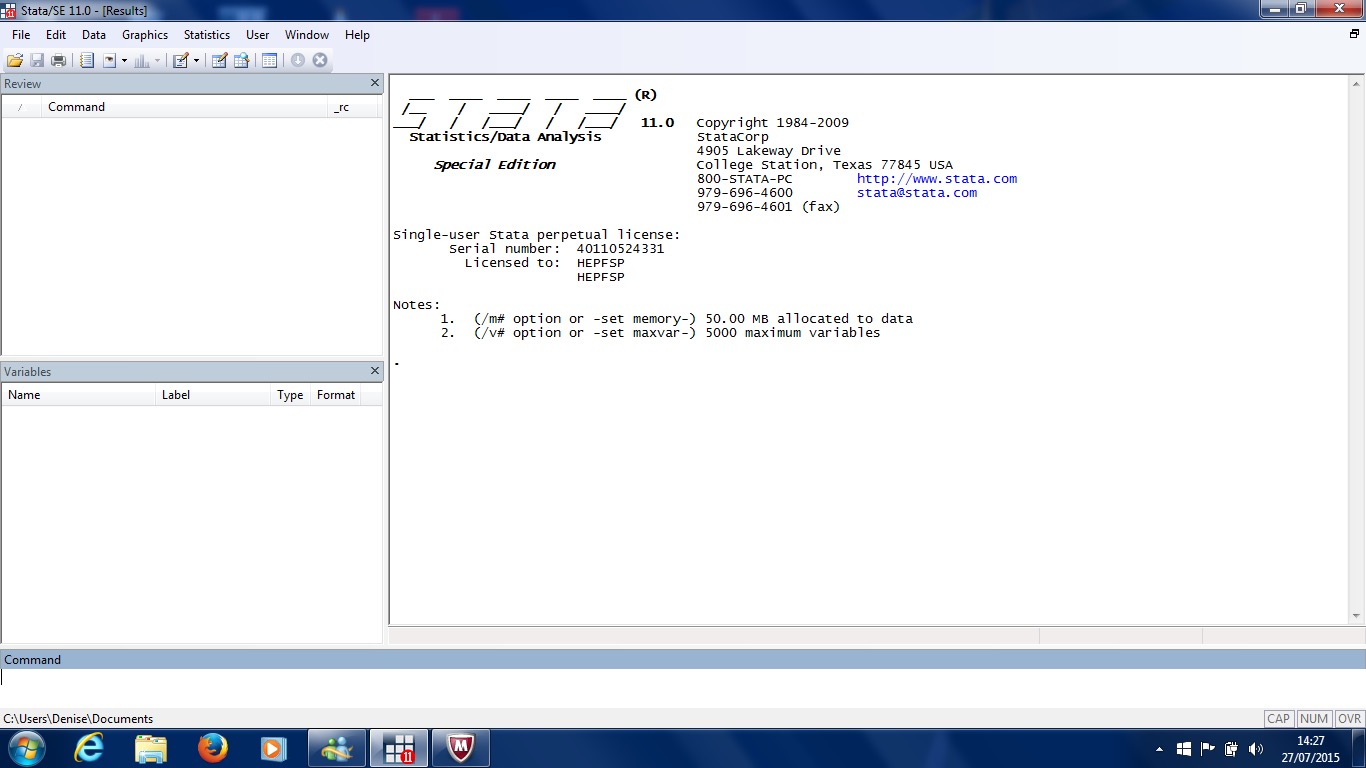
É um programa comercial para análise estatística de dados. Vamos utilizar as versões 10 e 13 para Windows, indistintamente com exceção para arquivos de banco de dados. Se estes forem criados na versão 13, deverão ser salvos também na 10 para permitir seu uso nesta versão.

O Stata funciona da seguinte forma:

1. Abre-se um banco de dados;
2. Escreve-se o comando desejado na janela de comando ou opta-se pela utilização do menu representado em caixas ou ícones apropriados, posicionados no topo da tela;
3. O comando será executado e os resultados serão apresentados na tela maior.

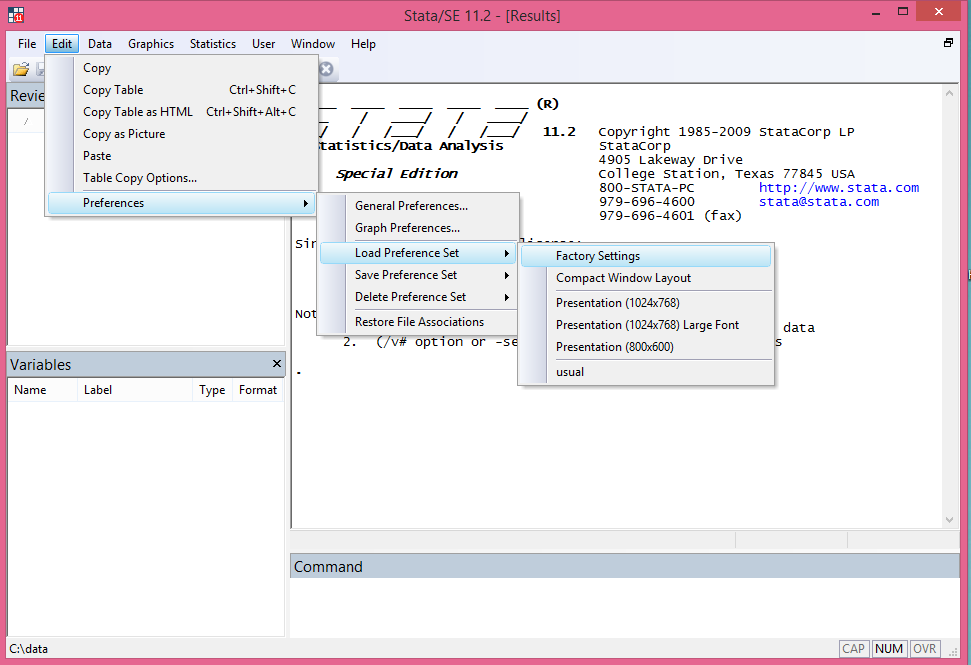
Informações importantes:

* O Stata trabalha com somente um banco de dados de cada vez;
* É possível salvar a sessão de trabalho em um arquivo do tipo texto;
* É possível escrever os comandos em modo de programa;
* O Stata é sensível para letras maiúsculas ou minúsculas.

*Janela de abertura*

* Janela Review: armazena os comandos já executados
* Janela Command: é onde os comandos são digitados
* Janela Variables: armazena o nome das variáveis do banco de dados
* Janela principal: armazena os resultados da execução do comando

Se uma das janelas da tela de abertura for fechada inadvertidamente, pode-se restaurar o layout original pelo menu Edit.



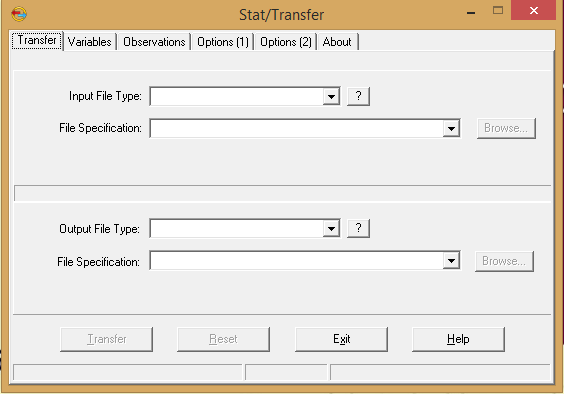
O Stata trabalha com arquivos de sete tipos de extensão

|  |  |
| --- | --- |
| Arquivo | Extensão |
| Executável | .exe |
| Que contém dados | .dta |
| Que contém comandos utilizados na sessão de trabalho | .log ou .smcl |
| Que contém uma lista de comandos | .do |
| Que contém sub-rotinas | .ado |
| Que contém gráficos | .gph |
| Que guarda a construção de gráficos | .grec |

Pode-se converter os dados em formato xls para dta utilizando o programa *Transfer*.

Utilizando o StatTransfer

****Localizar o programa por meio do menu do computador ou ícone na área de trabalho. Ao abri-lo aparecerá a janela abaixo:

****

*Para utilizar o transfer:*

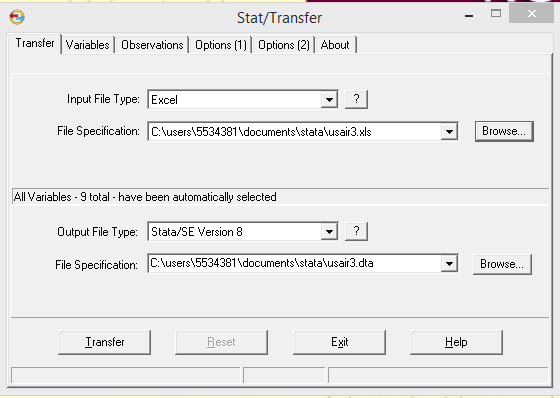
- No campo input file type: escolha o tipo de arquivo que quer converter para o programa Stata

- No campo file specification: por meio do browse selecione o arquivo que deseja converter

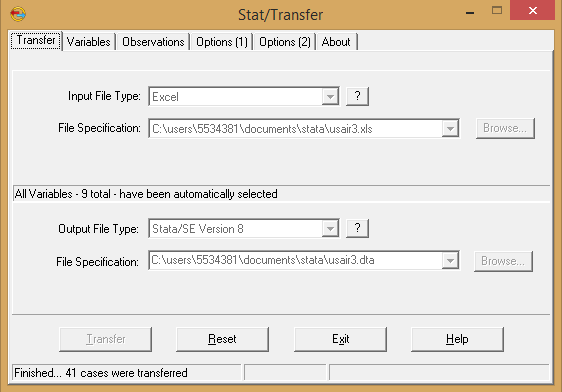
- No campo output file: escolha o programa para o qual deseja converter, por exemplo Stata

- No campo file specification: por meio do browse selecione o local e indique um nome para salvar o novo arquivo

Exemplo: Transferindo o banco de dados <usair> do formato Excel (.xls) para o formato Stata (.dta)



Com tudo checado, pode-se clicar no botão Transfer. Observe que aparecerá uma mensagem ao final da janela com o número de casos transferidos para o novo arquivo.



Para verificar se a transferência foi realizada com sucesso abra o arquivo no Stata.

Trazendo um banco de dados do Excel

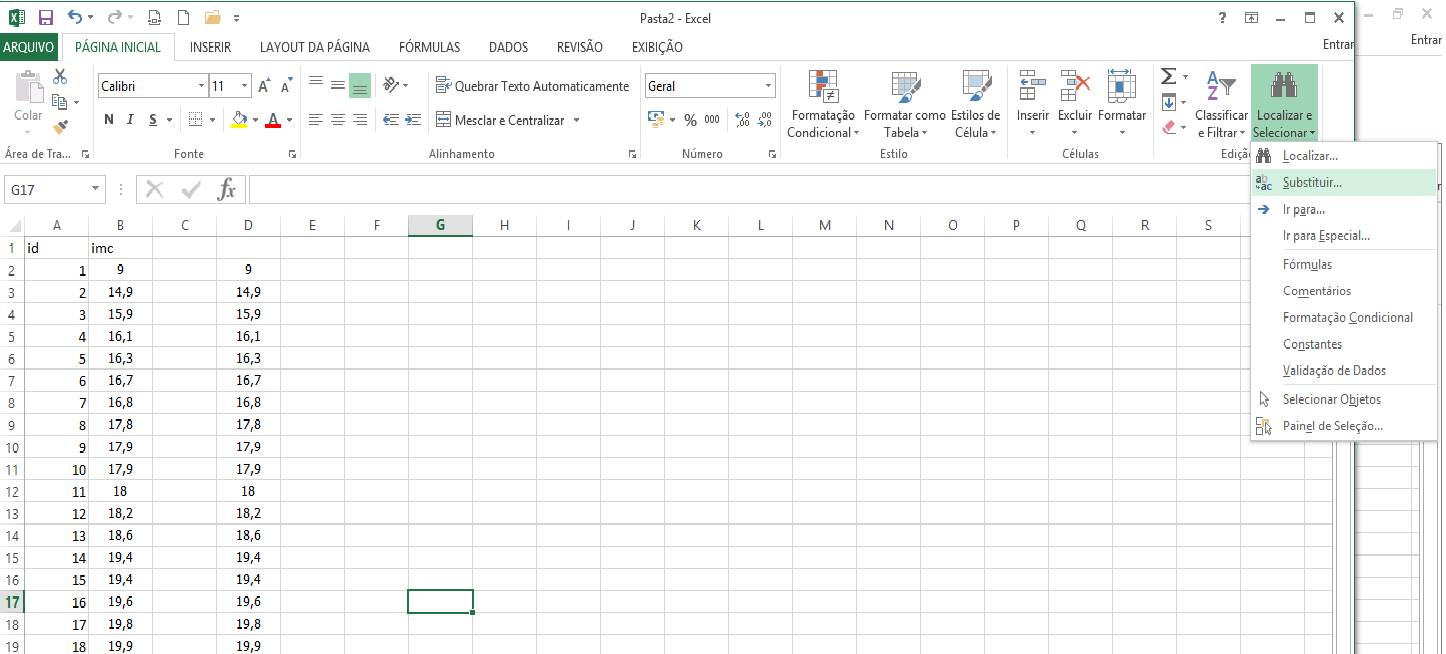
Também é possível trazer dados de um arquivo do Excel diretamente para o Stata, porém, é preciso que sejam trazidos somente os dados, mas não o nome das variáveis e as vírgulas das casas decimais devem ser substituídas por ponto.

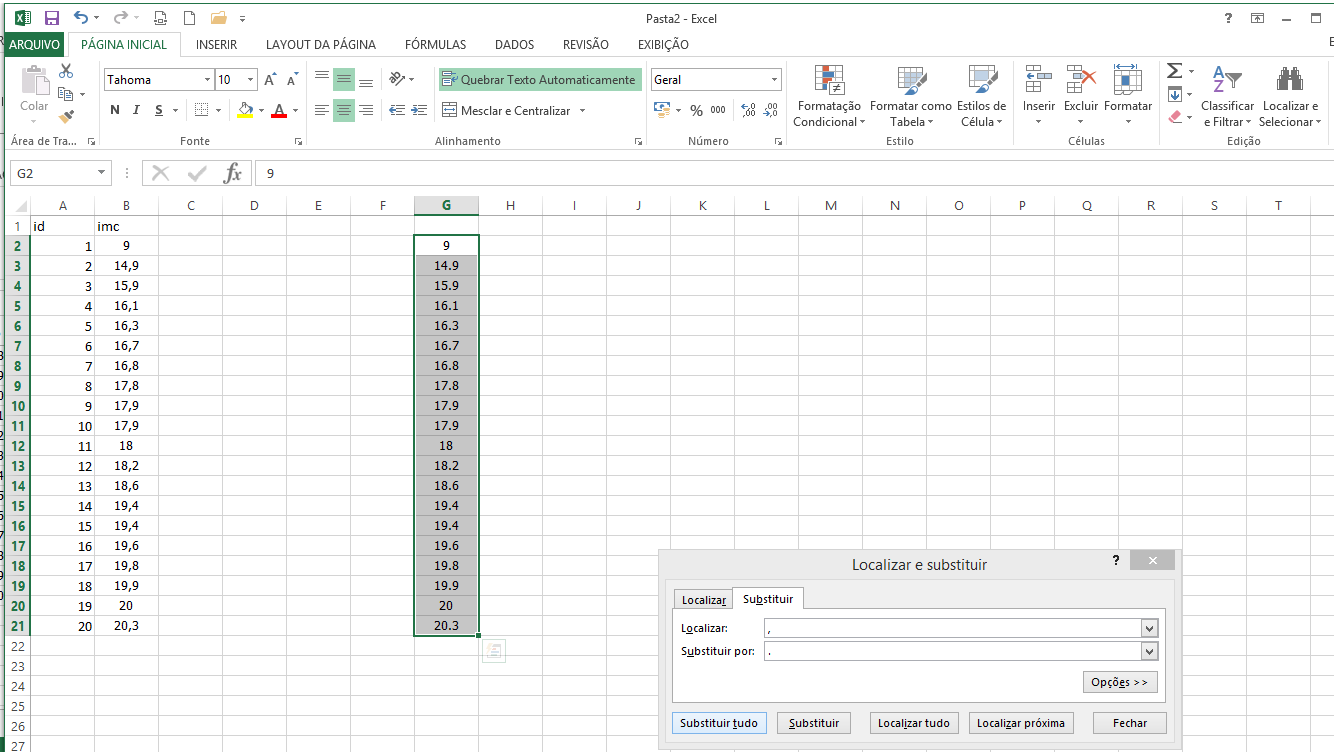
Exemplo

Utilizando-se os dados de imc de 20 pessoas:

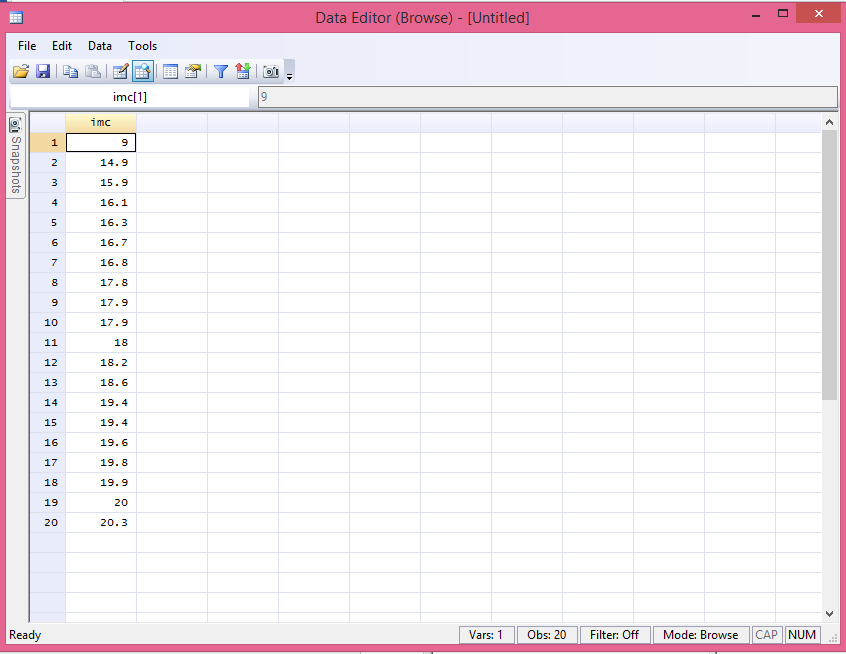
|  |  |
| --- | --- |
| id | imc |
| 1 | 9 |
| 2 | 14,9 |
| 3 | 15,9 |
| 4 | 16,1 |
| 5 | 16,3 |
| 6 | 16,7 |
| 7 | 16,8 |
| 8 | 17,8 |
| 9 | 17,9 |
| 10 | 17,9 |
| 11 | 18 |
| 12 | 18,2 |
| 13 | 18,6 |
| 14 | 19,4 |
| 15 | 19,4 |
| 16 | 19,6 |
| 17 | 19,8 |
| 18 | 19,9 |
| 19 | 20 |
| 20 | 20,3 |

Digitando-se no Excel com vírgula nas casas decimais. Copia-se os valores de imc para nova coluna e substitui-se a vírgula por ponto.



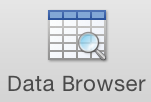
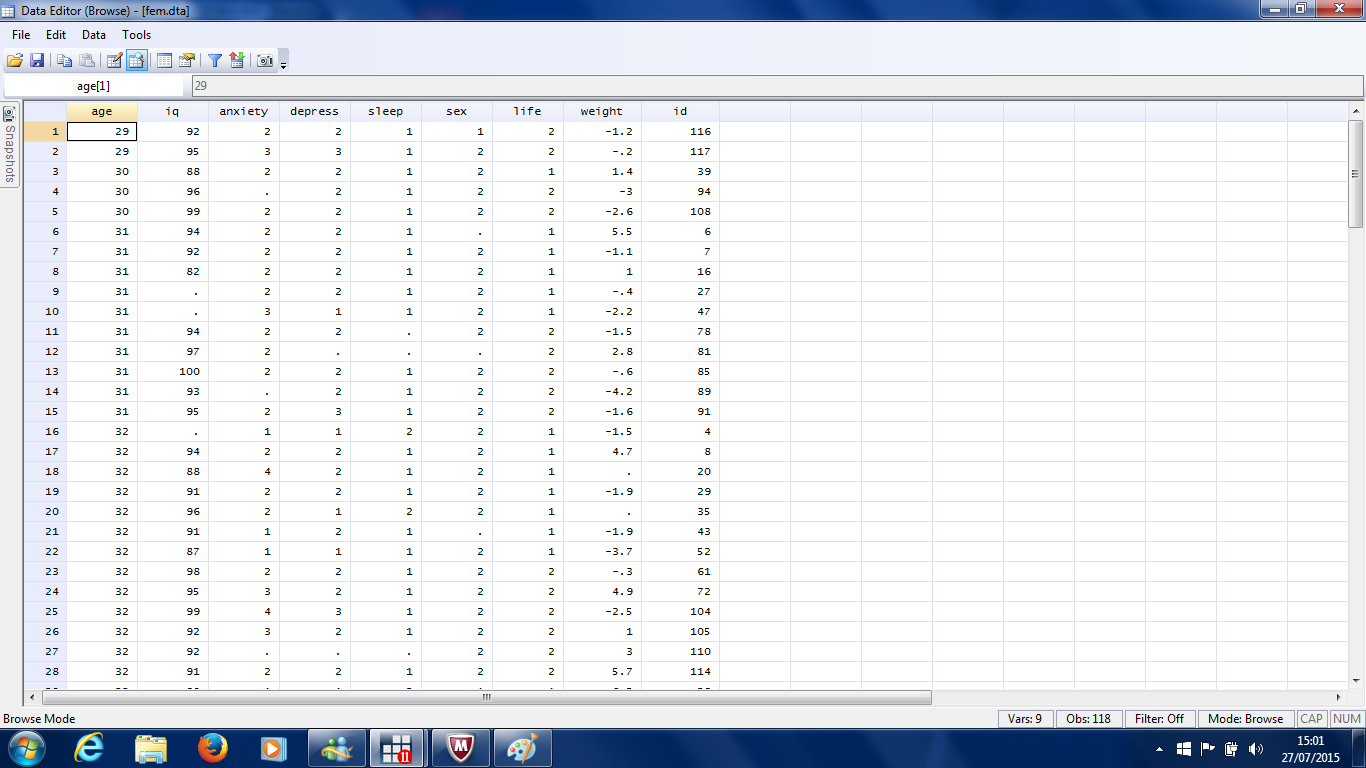


Copia-se a coluna de valores de imac (com ponto para casas decimais) e cola-se no ***data editor*** do Stata.



Abrindo um banco de dados no Stata

File>Open...> <usair3.dta>

Utilize o ícone Data Browse (sexto ícone, na terceira linha, com uma lupa) para olhar os dados contidos no banco de dados.

Notar que um banco de dados é construído em linhas e colunas formando campos numéricos, alfabéticos ou alfa numéricos, onde:

* Cada linha corresponde a um registro;
* Cada coluna corresponde a uma variável;
* Os nomes das variáveis são apresentados no cabeçalho;
* Os valores faltantes (missing values) são representados por pontos e tratados pelo Stata como valores muito grandes.

Operadores e expressões

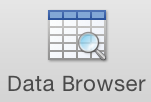
**Operadores**

|  |  |
| --- | --- |
| *Expressões* | *Significado* |
| < | Menor que |
| <= | Menor ou igual a |
| > | Maior que |
| >= | Maior ou igual a |
| = = | Igual a |
| ~= ou != | Diferente de |
| ~ | Não |
| & | e |
| | | ou |

**Expressões**

|  |  |
| --- | --- |
| *Expressões* | *Significado* |
| + | Soma |
| - | Subtração |
| \* | Multiplicação |
| / | Divisão |
| sqrt( ) | Raiz quadrada |
| exp( ) | Função exponencial |
| log() | Função logarítmica na base 10 |
| ln() | Função logarítmica na base e (logaritmo natural) |

Utilizando um banco de dados e abrindo um arquivo log

File>Open...> <fem.dta>

Utilize o ícone Data Browse para olhar os dados.

Abrindo um arquivo .log para armazenar a sessão de trabalho

Quarto ícone – log begin (optar por tipo log) <aula1.log>

Tudo o que for feito na sessão de trabalho, poderá ser guardado num arquivo log. Este arquivo é a documentação dos comandos usados (certos e errados) e saídas resultantes da execução destes. Pode ser editado em um editor de texto como por exemplo, o Word, se necessário.

Dica: Criar um arquivo log é útil como registro das análises realizadas

Comandos fundamentais

Utilizar os comandos e descrever o que eles fazem

**Quadro 1** – Preencha a coluna a respeito do que o comando faz.

|  |  |
| --- | --- |
| **Comando** | O que o comando faz? |
| ***describe ou desc*** |  |
|  |  |
|  |  |
| ***tab*** *age* |  |
|  |  |
| ***tab*** *iq* |  |
|  |  |
| ***tab*** *iq,* ***mis*** |  |
|  |  |
| ***tab1*** *anxiety depress sleep* |  |
|  |  |
|  |  |
| ***list*** *age* |  |
|  |  |
| ***help*** *sort* |  |
|  |  |
| ***sort*** *age* |  |
|  |  |
| ***gen******n****=\_n* |  |
| ***tab*** *n* |  |
|  |  |

Tipos de variáveis: string (caracteres, letras) e numérica. São armazenadas como registro de memória byte, int, long e float(variáveis numéricas) e str1 até str80(variáveis string de tamanhos diferentes).

Dando nomes para as categorias das variáveis

São necessários dois comandos

|  |  |
| --- | --- |
| Sintaxe dos comandos | Por exemplo, para a variável *anxiety*  (1 - não tem; 2 – leve; 3 – moderado; 4- grave) |
|  |  |
| ***label define*** *nome do label e rótulos* | ***label define*** *anx* 1 "não tem" 2 "leve" 3 "moderado" 4 "grave" |
|  |  |
| ***label val*** *nome da variável* *nome do label* | ***label val*** *anxiety anx* |
|  | ***tab*** *anxiety* |
|  | ***tab*** *anxiety,* ***nol*** |

**Exercício 1** - Faça o mesmo para as variáveis

tem depressão?*depress* (1 - não tem; 2 – leve; 3 – moderado; 4- grave)

pode dormir normalmente? *sleep* (1 – sim; 2 - não)

perdeu interesse em sexo? *sex* (1 – não; 2 – sim)

tem pensado recentemente em se matar? *life* (1 – não; 2- sim)

Veja no arquivo log o que você já fez clicando sobre o quinto ícone Log View e escolha

“view snapshot of log file”

Note as outras opções deste recurso

Salvar o banco com os rótulos das variáveis usando o menu

File> save as .... O novo banco deve ser nomeado como <femrotulos.dta>

Fechar o arquivo log: ***log close***

Fechar o banco de dados: ***clear***

Fazendo o resumo das variáveis contínuas

Abra novamente o bando de dados File> Open .... <femrotulos.dta>

Abra novamente um arquivo .log <aula1a.log>

**Quadro 2** - Preencha a coluna a respeito do que o comando faz.

|  |  |
| --- | --- |
| **Comando** | O que o comando faz? |
| ***help summarize*** |  |
|  |  |
|  |  |
| ***sum*** *age* |  |
|  |  |
| ***sum*** *age,****d*** |  |
|  |  |
| ***table*** *anxiety,* ***c(mean*** *age)* |  |
|  |  |
| ***sum*** *age iq weight* |  |
|  |  |
| ***bysort*** *anxiety:* ***sum*** *age* |  |
|  |  |

A apresentação de uma distribuição de frequências de uma variável contínua não é útil se forem considerados os valores individuais, por exemplo, veja a saída do comando ***tab*** *age.*

Tais variáveis podem ser apresentadas em intervalos de classe ou em categorias.

Apresentando uma variável contínua em intervalos de classe com amplitude cinco (anos).

***gen*** *age2=age*

***recode*** *age2* 25/29=1 30/34=2 35/39=3 40/44=4 45/49=5

***list*** *age age2*

***tab*** *age2*

***label define*** *age2* 1 “25 – 29” 2 “30 – 34” 3 “35 – 39” 4 “40 – 44” 5 “45 – 50”

***label val*** *age2 age2*

***tab*** *age2*

Criando categorias para uma variável contínua

Criando uma nova variável idade com pontos de corte dados pelos percentis

Por exemplo dividindo-se a distribuição em cinco partes

***centile*** *ag*e,***centile(20)***

***centile*** *age*,***centile(40)***

***centile*** *age*,***centile(60)***

***centile*** *age*,***centil*e(80)**

***gen*** *age3=age*

***recod***e *age3* 25/32=1 33/35=2 36/40=3 41/42=4 43/50=5

***list*** *age age3*

***tab*** *age3*

***label define*** *age3* 1 “25 – 32” 2 “33 – 35” 3 “36 – 40” 4 “41 – 42” 5 “43 – 50”

***label val*** *age3 age3*

***tab*** *age3*

Fazendo tabelas cruzadas

**Quadro 3** - Preencha a coluna a respeito do que o comando faz.

|  |  |
| --- | --- |
| **Comando** | O que o comando faz? |
| ***tab*** *depress anxiety* |  |
|  |  |
|  |  |
| ***tab*** *depress anxiety,****row*** |  |
|  |  |
|  |  |
| ***tab*** *depress anxiety,****col*** |  |
|  |  |
|  |  |

**Exercício 2**

1. Utilizando o arquivo <fem.dta>, construa uma nova variável *age4* e crie duas categorias com ponto de corte na média. Inclua o valor da média na primeira categoria.

***sum*** *age*

***gen*** *age4=age*

***recode*** *age4* 29/37=1 38/46=2

***label define*** *age4* 1 “29 – 37” 2 “38 – 46”

***label val*** *age4 age4*

***tab*** *age4*

1. Faça uma nova variável *iq* em intervalos de classe de tamanho 5.
2. Faça uma nova variável *iq* com duas categorias. Utilize o valor da mediana como ponto de corte.
3. Feche o arquivo log utilizando o comando **log close**.
4. Abra o arquivo log no Word para edição

**Exercício 3**

1. Abra o arquivo <Botucatu.dta>
2. Abra um arquivo com extensão .log para armazenar a sessão de trabalho
3. Atribua um rótulo para as categorias da variável *sexo* (1-masculino; 2-feminino)
4. Faça uma tabela de cada variável

***tab*** *sexo*

***tab*** *pesonasc*

1. Resuma a variável peso ao nascer, utilizando a opção ***details***

***sum*** *pesonasc****,d***

Qual é o peso ao nascer médio? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Qual é o peso ao nascer mediano? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Qual é o valor mínimo? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Qual é o valor máximo? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Qual é o valor do peso no percentil 25? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Qual é o valor do peso no percentil 50? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Qual é o valor do peso no percentil 75? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Apresente a variável peso ao nascer em intervalos de classe de tamanho 500g

**gen** *pesonasc2***=***pesonasc*

**recode** *pesonasc2* 500/999=1 1000/1499=2 1500/1999=3 2000/2499=4 2500/2999=5 3000/3499=6 3500/3999=7 4000/4499=8

**tab** *pesonasc2*

**label define** *pesonasc2* 1 “500 – 999” 2 “1000 – 1499” 3 “1500 – 1999” 4 “2000 – 2499” 5 “2500 – 2999” 6 “3000 – 3499” 7 “3500 – 3999” 8 “4000 – 4499”

**label val** *pesonasc2 pesonasc2*

**tab** *pesonasc2*

1. Construa uma nova variável peso ao nascer em duas categorias, acima e abaixo de 2500 g. Inclua o valor 2500g na segunda categoria.

**gen** *pesonasc3*=*pesonasc*

**recode** *pesonasc3* **min**/2499=1 2500/**max**=2

**tab** *pesonasc3*

**label define** *pesonasc3* 1 “baixo peso” 2 “não baixo peso”

**label val** *pesonasc3 pesonasc3*

**tab** *pesonasc3*

1. Faça uma tabela cruzando as variáveis *pesonasc3* e *sexo*. Opte por colocar o 100% na variável *sexo.*

Criando um banco de dados

Considere os dados a seguir:

São apresentados dados sobre poluição atmosférica de 41 cidades nos Estados Unidos, coletados por Sokal e Rohlf (1981). A variável dependente é a concentração anual de dióxido de enxofre (SO2) em microgramas por metro cúbico. Os dados de SO2 correspondem à média dos anos 1969 a 1971. Os valores de 6 variáveis explanatórias são também apresentados, duas das quais relacionadas a ecologia humana e quatro com o clima. (Sophia Rabe-Hesketh; Brian S Everitt, 2007)

|  |  |  |
| --- | --- | --- |
| Nome da variável | Descrição da varável | Nome da variável no banco de dados |
| Cidade | cidade onde os dados foram coletados | *cidade2* |
| id | número de identificação da cidade | *id* |
| SO2 | dióxido de enxofre | *so22* |
| Temp | temperatura média anual em 0F (Fahrenheit) | *temp2* |
| Manufatura | número de fábricas com 20 ou mais trabalhadores | *manufa2* |
| Pop | tamanho da população (censo de 1970) em milhares | *pop2* |
| Vento | velocidade média anual em milhas por hora | *vento2* |
| Precipitação | precipitação média anual (polegadas) | *precipita2* |
| Dias | número médio anual de dias de precipitação | *dias2* |

Dados do arquivo <usair.dta>

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *cidade2* | *id* | *so22* | *temp2* | *manufa2* | *pop2* | *vento2* | *precipita2* | *dias2* |
| Phoenix | 1 | 10 | 70.3 | 213 | 582 | 6.0 | 7.05 | 36 |
| Little Rock | 2 | 13 | 61.0 | 91 | 132 | 8.2 | 48.52 | 100 |
| San Francisco | 3 | 12 | 56.7 | 453 | 716 | 8.7 | 20.06 | 67 |
| Denver | 4 | 17 | 51.9 | 454 | 515 | 9.0 | 12.95 | 86 |
| Hartford | 5 | 56 | 49.1 | 412 | 158 | 9.0 | 43.37 | 127 |
| Wilmington | 6 | 36 | 54.0 | 80 | 80 | 9.0 | 40.25 | 114 |
| Washington | 7 | 29 | 57.3 | 434 | 757 | 9.3 | 38.89 | 111 |
| Jackson | 8 | 14 | 68.4 | 136 | 529 | 8.8 | 54.47 | 116 |
| Miami | 9 | 10 | 75.5 | 207 | 335 | 9.0 | 59.80 | 128 |
| Atlanta | 10 | 24 | 61.5 | 368 | 497 | 9.1 | 48.34 | 115 |
| Chicago | 11 | 110 | 50.6 | 3344 | 3369 | 10.4 | 34.44 | 122 |
| Indianapolis | 12 | 28 | 52.3 | 361 | 746 | 9.7 | 38.74 | 121 |
| Des Moines | 13 | 17 | 49.0 | 104 | 201 | 11.2 | 30.85 | 103 |
| Wichita | 14 | 8 | 56.6 | 125 | 277 | 12.7 | 30.85 | 82 |
| Louisville | 15 | 30 | 55.6 | 291 | 593 | 8.3 | 43.11 | 123 |
| New Orleans | 16 | 9 | 68.3 | 204 | 361 | 8.4 | 56.77 | 113 |
| Baltimore | 17 | 47 | 55.0 | 625 | 905 | 9.6 | 41.31 | 111 |
| Detroit | 18 | 35 | 49.9 | 1064 | 1513 | 10.1 | 30.96 | 129 |
| Minneapolis | 19 | 29 | 43.5 | 699 | 744 | 10.6 | 25.94 | 137 |
| Kansas | 20 | 14 | 54.5 | 381 | 507 | 10.0 | 37.00 | 99 |
| St Louis | 21 | 56 | 55.9 | 775 | 622 | 9.5 | 35.89 | 105 |
| Omaha | 22 | 14 | 51.5 | 181 | 347 | 10.9 | 30.18 | 98 |
| Albuquerque | 23 | 11 | 56.8 | 46 | 244 | 8.9 | 7.77 | 58 |
| Albany | 24 | 46 | 47.6 | 44 | 116 | 8.8 | 33.36 | 135 |
| Buffalo | 25 | 11 | 47.1 | 391 | 463 | 12.4 | 36.11 | 166 |
| Cincinnati | 26 | 23 | 54.0 | 462 | 453 | 7.1 | 39.04 | 132 |
| Cleveland | 27 | 65 | 49.7 | 1007 | 751 | 10.9 | 34.99 | 155 |
| Columbia | 28 | 26 | 51.5 | 266 | 540 | 8.6 | 37.01 | 134 |
| Philadelphia | 29 | 69 | 54.6 | 1692 | 1950 | 9.6 | 39.93 | 115 |
| Pittsburgh | 30 | 61 | 50.4 | 347 | 520 | 9.4 | 36.22 | 147 |
| Providence | 31 | 94 | 50.0 | 343 | 179 | 10.6 | 42.75 | 125 |
| Memphis | 32 | 10 | 61.6 | 337 | 624 | 9.2 | 49.10 | 105 |
| Nashville | 33 | 18 | 59.4 | 275 | 448 | 7.9 | 46.00 | 119 |
| Dallas | 34 | 9 | 66.2 | 641 | 844 | 10.9 | 35.94 | 78 |
| Houston | 35 | 10 | 68.9 | 721 | 1233 | 10.8 | 48.19 | 103 |
| Salt Lake City | 36 | 28 | 51.0 | 137 | 176 | 8.7 | 15.17 | 89 |
| Norfolk | 37 | 31 | 59.3 | 96 | 308 | 10.6 | 44.68 | 116 |
| Richmond | 38 | 26 | 57.8 | 197 | 299 | 7.6 | 42.59 | 115 |
| Seattle | 39 | 29 | 51.1 | 379 | 531 | 9.4 | 38.79 | 164 |
| Charleston | 40 | 31 | 55.2 | 35 | 71 | 6.5 | 40.75 | 148 |
| Milwaukee | 41 | 16 | 45.7 | 569 | 717 | 11.8 | 29.07 | 123 |

Para criar um banco de dados no Stata clique no oitavo ícone (Data Editor) na barra de ferramentas no topo da tela.

Data Editor

* Digite os dados apenas. Os nomes das variáveis serão atribuídos posteriormente;
* Digite o nome das cidades com acentuação podendo ser duas ou mais palavras. Neste caso não é necessário deixá-las entre aspas;
* Para as variáveis numéricas deve-se utilizar o ponto para as casas decimais;
* Para atribuir nomes às variáveis clique com o mouse do lado direito sobre o nome da variável atribuído pelo Stata (var1) e escolha a opção variable properties. Na caixa name digite o nome da variável (cidade). Faça o mesmo para todas as variáveis;
* Salve o arquivo de dados com o nome <usair2.dta*>*.

Validação da digitação

É recomendável validar a digitação para todo banco de dados após sua criação. Isto é feito por meio de dupla digitação. Uma das possibilidades é criar dois bancos de dados e compará-los após a digitação. Como no Stata não tem esta sub rotina, vamos fazer a comparação de modo artesanal.

Iremos fornecer um segundo banco de dados: usair1.dta para ser comparado com o banco usair2.dta digitado pelo aluno.

Para tanto abre-se o banco usair2.dta e utiliza-se os comandos:

***merge******1:1*** *id* ***using***<usair1.dta>

***gen*** *difso2=so22-so2*

***tab*** *difso2*

***gen*** *diftemp=temp2-temp*

***tab*** *diftemp*

***gen*** *difmanuf=manufa2-manufa*

***tab*** *difmanufa*

***gen*** *difpop=pop2-pop*

***tab*** *difpop*

***gen*** *difvento=vento2-vento*

***tab*** *difvento*

***gen*** *difprecipita=precipita2-precipita*

***tab*** *difprecipita*

***gen*** *difdias=dias2-dias*

***tab*** *difdias*

Salvar o banco criado pelo comando merge utilizando o Menu🡪File🡪 Save as <usairmerge.dta>

Se os bancos estiverem iguais, então todos os valores das variáveis com início dif serão iguais a zero. Se alguma delas apresentar um valor diferente de zero então será necessário verificar onde está o erro e corrigi-lo. No final do processo os dois bancos deverão estar iguais**.**

Para corrigir localize a variável que contém valores diferentes por meio da variável que contém a diferenças entre a primeira e segunda digitação. Para checar, utilize o comando por exemplo para a variável dias:

***list*** *id dias2 dias difdias*

Uma vez que já se sabe qual é o dado que tem que ser corrigido, cheque na tabela (ou no questionário) o valor correto e identifique em qual dos bancos de dados está o erro.

Utilizando os bancos originais faça a correção utilizando o comando abaixo, para que ambos fiquem idênticos. Você pode utilizar o Data Editor ou substituir o valor errado utilizando o comando:

**replace** *dias*==<x> **if** *id*==<y>

***save, replace*** (esse comando salva o banco que está sendo utilizado substituindo o antigo banco pelo novo)

Após corrigir todas as discordâncias e obter dois bancos de dados iguais, pode-se descartar o banco criado pelo comando merge.

Análise exploratória

Atenção: Os gráficos deverão ser salvos um a um.

Para salvar o gráfico utilize a caixa de apresentação do gráfico: File> Save as...<escolha extensão emf>

***- Box plot***

**graph box** *so2*

**graph box** *so2*, **ylabel(0(10)110)**

**tab** *so2*

**list** *id so2* **if** *so2>65*



Identificando as cidades que apresentam concentrações tão acima das demais (outliers)

**list** *cidade so2* ***if id==***11 ***|******id==***29 ***| id==***31

Faça o box plot para as demais variáveis e investigue a existência de valores aberrantes

Variável *temperatura*

**graph box** *temp*

**tab** *temp*

**list** *id temp* **if** *temp>72*

**list** *cidade temp* ***if*** *id==9*

Quais são as cidades que apresentam temperatura tão acima das demais?

Variável *manufatura*

***graph box*** *manufa*

***tab*** *manufa*

***lis****t id manufa* ***if*** *manufa>1000*

***list*** *cidade manufa* ***if*** *id==11| id==18 | id==27| id==29*

Quais são as cidades que apresentam número de fábricas tão acima que as demais?

Variável *população*

***graph box*** *pop*

***tab*** *pop*

***list*** *id pop* ***if*** *pop>1500*

***list*** *cidade pop* ***if*** *pop>1500*

Quais são as cidades que apresentam número de habitantes tão acima que as demais?

Variável *vento*

***graph box*** *vento*

O que podemos dizer do gráfico da variável vento?

Variável precipitação

***graph box*** *precipita*

***tab*** *precipita*

***list*** *id cidade precipita* ***if*** *precipita<10*

Quais são as cidades que apresentam precipitação média anual tão menor que as demais?

Variável dias

***graph box*** *dias*

***tab*** *dias*

***list*** *id cidade dias* ***if*** *dias>165 | dias<60*

Quais são as cidades que apresentam número de dias de precipitação tão abaixo ou acima das demais?

Para verificar valor ***outlier*** considerando todas as variáveis conjuntamente

***hadimvo*** *so2 temp manufa pop vento precipita dias,* ***gen(****odd****)***

***list*** *id cidade so2 temp manufa pop vento precipita dias odd* ***if*** *odd==1*

Histograma (variáveis quantitativas – discreto ou contínuo)

**Quadro 4** - Teste os seguintes comandos e anote o que cada um faz

|  |  |
| --- | --- |
| **Comando** | O que o comando faz? |
| ***histogram*** *so2* |  |
| ***histogram*** *so2****, xlabel****(10(10)110)* ***bin****(10)* ***xtitle*** *(“SO2 microgramas por m3”)* |  |
| ***histogram*** *so2,****xlabel****(10(10)110)* ***bin****(10)* ***freq*** |  |
| ***histogram*** *so2,* ***xlabel****(10(10)110)* ***bin****(10)* ***freq norm*** |  |

Se for necessário verificar normalidade da distribuição

**Quadro 5 -** Teste os seguintes comandos e anote o que cada um faz

|  |  |
| --- | --- |
| **Comando** | O que o comando faz? |
| ***ladder*** *so2* |  |
| ***gladde****r so2* |  |

Faça o histograma para as variáveis temperatura, vento, precipitação e dias de precipitação. Investigue se as variáveis têm distribuição normal. Se a variável não apresentar distribuição normal. Qual transformação dos dados poderia normaliza-lo?

**Quadro 6** - Utilize o comando gladder e identifique as transformações que normalizam os dados.

|  |  |
| --- | --- |
| **Variável** | Transformações |
| *Temp* |  |
| *Vento* |  |
| *Precipitação* |  |
| *Dias* |  |

Criar uma variável nova para classificar manufatura em quatro classes: 1 a 49; 50 a 299; 300 a 999; 1000 a 3500.

***gen*** *manufa1=manufa*

***recode*** *manufa1 1/49=1 50/299=2 300/999=3 1000/3500=4*

***tab*** *manufa1*

***label define*** *manufa1 1 "1 – 49" 2 "50 – 299" 3 "300 – 999" 4 "1000 – 3500"*

***label val*** *manufa1 manufa1*

***tab*** *manufa1*

Investigue o valor médio do dióxido de enxofre segundo as categorias de manufatura

***bysort*** *manufa1:****sum*** *so2*

Investigando a correlação entre as variáveis

* ***Gráfico de dispersão***

***twoway (scatter*** *so2 temp),* ***xtitle****(Temperatura(F))* ***ytitle****(SO2)*

**

***twoway (scatter*** *so2 temp,* ***mlabel****(cidade)*

Verificando a normalidade

***ladder*** *so2*

***ladder*** *temp*

Gráfico de dispersão com as variáveis transformadas em logaritmo

***gen*** *logso2=log(so2)*

***gen*** *logtemp=log(temp)*

***twoway*** *(scatter logtemp logso2)*

Calculando os coeficientes de correlação:

***cor*** *logso2 logtemp*

O que faz o comando? ***pwcorr*** *logso2 logtemp,* ***sig obs***

Outro coeficiente de correlação

***spearman*** *so2 temp*

O diagrama de dispersão sugere relação linear entre as variáveis?

Você pode usar o coeficiente de correlação linear de Pearson para investigar a correlação entre as variáveis *so2* e *temp*?

Faça o mesmo para

*so2 X pop*

*so2 X vento*

*so2 X precipita*

*so2 X dias*

*so2 X manufatura*

***- Gráficos apresentados como uma matriz de diagramas de dispersão***

***graph matrix*** *so2 temp manufa pop vento precipita dias*



Notar que o gráfico na linha 1, coluna 2 apresenta no eixo Y a variável so2 e no eixo X, a variável temp. O gráfico posicionado na linha 2, coluna 1 apresenta no eixo Y a variável temp e no eixo X a variável so2.

Para quais variáveis seria interessante calcular o coeficiente de correlação linear de Pearson?

Crie uma variável nova para so2 de tal forma que apresente distribuição normal

***gen*** *logso2=****log****(so2)*

***twoway (scatter*** *logso2 temp)*

Compare este gráfico com aquele criado para as variáveis so2 e temp. Qual a diferença entre eles?

O coeficiente de correlação entre log(so2) e temp é diferente do coeficiente para so2 e temp?

Gráfico de barras

Abrir o banco <fem.dta>

***graph bar (count)*** *sleep*

Notar que a variável sleep tem duas categorias mas o comando não as representa separadamente. Para fazer o gráfico de barras considerando a frequência de ocorrência de cada uma é necessário gerar uma variável que funcione como contador, no caso será denominada freq

***gen*** *freq=1*

***graph bar (count)*** *freq,* ***over****(sleep)*

***graph bar (count****) freq,* ***over****(sleep,* ***relabel****(1 "Sim" 2"Não"))* ***ytitle****(Frequência)*

**graph bar (count) freq, over(sleep, relabel(1 `" "Com problemas" "de sono" "' 2 `" "Sem problemas" "de sono" "')) ytitle(Frequência)**

******

***graph bar (count)*** *freq****, over****(sleep****, relabel****(1 "Sono\_Normal" 2"Sono\_alterado"))* ***over****(sex,* ***relabel****(1 "Não\_perdeu" 2"Perdeu"))* ***ytitle****(Frequência)*



O Stata, nos gráficos de barras, apresenta a representação de estatísticas tais como soma, média, desvio padrão, segundo as categorias de variáveis. Neste caso é necessário que a variável seja quantitativa.

***graph bar (mean)*** *weight,* ***over*** *(sleep,* ***relabel****(1 "Sono\_Normal" 2"Sono\_alterado"))* ***ytitle****(Média de perda de peso)*



Construindo um gráfico de barras com duas variáveis

Abrir o banco <usair.dta>

Criar um gráfico de barras para a variável manufa1:

***gen*** *freq=1*

***graph bar (count)*** *freq****, over****(manufa)* ***stack***

Criando a variável so2 em duas categorias (abaixo e acima da média)

***gen*** *so2g=so2*

***recode*** *so2g 8/30=1 31/110=2*

***tab*** *so2g*

***label define*** *so2g 1 "8 – 30" 2 "31 – 110"*

***label val*** *so2g so2g*

***tab*** *so2g*

***graph bar (count)*** *freq,* ***over****(so2g)* ***over****(manufa1)* ***stack bar****(1,color(green))* ***ytitle*** *(Frequência)*

**

***Diagrama*** linear

Utilize o banco de dados <earningratios.dta>

*The Open University (1993) MDST242 Statistics in Society Unit A2: Earnings. 2nd edition. Milton Keynes: The Open University. Table 1.4. (Hand DJ et al. A handbook of small data sets pg 49)*

Esses dados foram coletados como parte de uma pesquisa britânica sobre a igualdade de salário entre homens e mulheres. São originários de dados do governo britânico – Pesquisa Anual de Salários – que utiliza a distribuição de salários de homens e mulheres empregados em tempo integral em relação às taxas de remuneração dos adultos. A pesquisa informou a mediana, os quartis e os decis mais alto e mais baixo por ganho semanal. As taxas de ganho fornecem tanto quantis para mulheres como porcentagem em relação aos homens. Os dados são apresentados com intervalos de 2 anos entre 1970 e 1990. Juntamente com os dados de 1975, ano em que a “Ação para Igualdade de Pagamentos de 1970” foi implementado inteiramente pela primeira vez. Taxa de ganho F/M (%):

***twoway (line*** *decil10 quartil4 median quartil1 decil1 ano)*

***twoway connected*** *decil10 quartil4 median quartil1 decil1 ano,* ***msymbol (T O S X D)******color (black gs3 gs7 gs11 gs13) ytitle****(Frequência)*

**

Atenção: para ver as opções de símbolo utilizar o help ***symbolstyle*** e para as opções de cor help ***colorstyle***

Outra opção modificando os símbolos, cor e estilo das linhas:

***twoway connected*** *decil10 quartil4 median quartil1 decil1 ano,* ***msymbol(T O S X D)******clpattern(shortdash shortdash\_dot solid shortdash\_dot shortdash)******color (cranberry gs8 gs0 gs8 cranberry)***

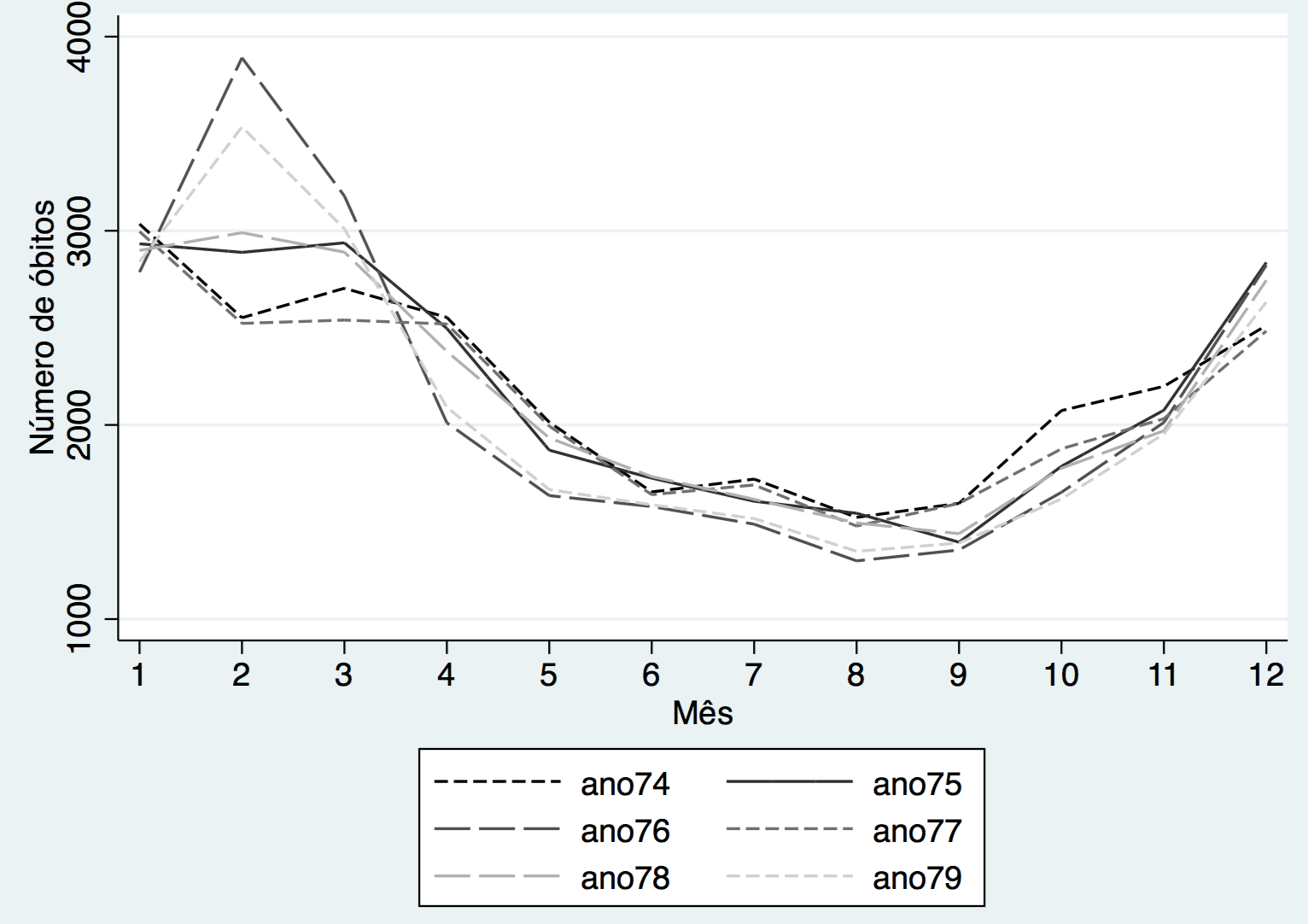
Utilize o banco <lungdeaths.dta>

*Diggle. PJ (1990). Time Series: A Biostatistical Introduction. Oxford: Oxford University Press. Table A.3 combining both sexes.*

Esse banco de dados traz informações do número de óbitos mensais por bronquite, enfisema e asma no Reino Unido entre 1974 a 1979 para ambos os sexos.

Podemos dizer que o número de mortes depende do período do ano? Alguma coisa mudou entre 1974 e 1979?

***twoway (line*** *ano74 ano75 ano76 ano77 ano78 ano79 mes,* ***lpattern (- l \_- \_--\_# \_-- -.)******color (black gs3 gs5 gs7 gs11 gs13))****,* ***xlabel(1(1)12) xtitle****(Mês)* ***ytitle****(Número de óbitos)*



Comandos mais usados

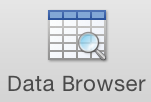
***Substituindo valores***

Utilize o banco de dados <femmissing.dta**>**

Abra um arquivo log com o nome <aula3.log>

Utilize os ícones

**Data Editor –** para edição de dados



**Browse** – Para olhar o banco de dados

Se o interesse for apenas olhar os valores, então a melhor opção é o Browse porque o Data Editor permite modificar os dados e, apesar do Stata fazer uma boa documentação, é possível alterar o banco de dados sem querer e sem perceber e isto causaria um problema importante.

Utilize o comando desc para conhecer as variáveis

Note que para algumas variáveis os valores faltantes (missing) são codificados como -99. Nestes casos é necessário substitui-los por um ponto (.) para que o Stata dê o tratamento como missing value.

***tab*** *iq*

***gen*** *iq2=iq*

***replace*** *iq2=.* ***if*** *iq== -99*

***tab*** *iq2*

***list*** *iq2 iq*

***tab*** *iq2,****mis***

***drop*** *iq*

***ren*** *iq2 iq*

*(****rename*** *nome atual da variável nome novo da variável)*

***ou***

***tab*** *sex*

***mvdecode*** *sex,****mv****(-99)*

**Atenção:** Antes de começar uma análise é preciso verificar como os valores faltantes estão apresentados no banco de dados. Se os dados forem digitados em um outro programa como o Epi Info e os dados tiverem sido convertidos para o Stata, pelo programa Transfer, se originalmente os valores faltantes tiverem sido deixados em branco, na transferência, o Transfer substituirá o campo vazio por ponto (.).

***Outros comandos***

É possível utilizar melhor a área onde o Stata armazena os dados utilizando o comando

***compress***

Se for de interesse alterar a ordem das variáveis no banco, utilize o comando

***order*** *<lista de variáveis>*

Se for de interesse ordenar os valores de uma variável, utilize o comando

***sort*** *<nome da variável>*

Se for de interesse saber o valor de um campo por exemplo idade para o indivíduo que ocupa a posição 10, ou seja a linha 10, então utilize o comando

***list*** *age in 10*

Se for de interesse alterar a codificação de uma variável por exemplo variável sexo a categoria 1 deve ser 0 e a 2 deve ser 1.

***gen*** *sex2=sex*

***recode*** *sex2 1=0 2=1*

***tab*** *sex sex2*

Trabalhando com variáveis data

O Stata vê data como tempo decorrido desde 01 de janeiro de 1960 (elapsed data) ou %d que corresponde ao número de dias a partir de 01 de janeiro de 1960. Assim, 0 corresponde a 01 de janeiro de 1960; 2 corresponde a 02 de janeiro de 1960.

Crie um banco de dados com as seguintes datas de nascimento (variável datanasc).

|  |  |  |
| --- | --- | --- |
| 12/04/1955 | 15/07/1957 | 21/09/1971 |
| 14/03/1960 | 30/12/1980 | 16/01/1970 |
| 19/03/1964 | 30/03/1955 | 12/08/1999 |

Utilizar os comandos e descrever o que eles fazem

**Quadro 7** – Preencha o quadro descrevendo o que cada comando faz

|  |  |
| --- | --- |
| ***gen*** *datanum=****date****(datanasc,* ***“DMY#”)*** |  |
| ***gen*** *datastring=****string****(datanum,* ***“%td”)*** |  |
| ***gen*** *idade=(date(“20/08/2015”,”DMY”)-datanum)/365.25* |  |

Fechar o banco de dados: ***Clear***

Mudando a forma de apresentação dos dados

Mudando a apresentação do banco de formato wide para long

Abrir o arquivo <calorias1.dta>

Observe que cada pessoa foi observada em duas consultas, com aferição de calorias ingeridas. O banco contém as variáveis id, cal1, cal2 e sexo. Este formato é denominado no Stata wide. Para fazer uma só variável com valores de caloria e outra com a indicação da consulta (primeira ou segunda), é necessário mudar o formato do banco para long. Neste caso, utilizar o comando:

***reshape long*** *cal,* ***i****(id)* ***j****(dieta)*

Para fazer isto as variáveis originais (cal1, cal2) precisam ter o mesmo prefixo e o Stata vai denominar a nova variável somente o que está no prefixo (cal).

Se forem mais variáveis como *var1 var2*, é possível incluir

***reshape long*** *cal var1 var2...,* ***i****(id)* ***j****(dieta)*

Mudando a apresentação do banco de formato long para wide

***reshape wide*** *cal,* ***i****(id)* ***j****(dieta)*

***reshape wide*** *cal var,* ***i****(id)* ***j****(dieta)*

Combinando bancos de dados

***Posicionando um banco de dados em baixo de outro***

Este comando permite combinar dois bancos de dados sendo que um é posicionado no final do o outro. O comando ***append***somente combina bancos com a mesma estrutura, por exemplo o primeiro banco tem os 100 primeiros registros e o segundo, os 100 seguintes. Esquema:

<banco1.dta>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | var1 | var2 | var3 | var4 | var5 |
| 1 | x | x | x | x | x |
| 2 | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| 100 | x | x | x | x | x |

<banco2.dta>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | var1 | var2 | var3 | var4 | var5 |
| 101 | x | x | x | x | x |
| 102 | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| 200 | x | x | x | x | x |

Banco resultante da combinação dos dois anteriores

<banco12.dta>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | var1 | var2 | var3 | var4 | var5 |
| 1 | x | x | x | x | x |
| 2 | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| 100 | x | x | x | x | x |
| 101 | x | x | x | x | x |
| 102 | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| 200 | x | x | x | x | x |

Para fazer este tipo de combinação, utilizar os comandos:

***use*** *<banco1.dta>*

***append using*** *<banco2.dta>*

***desc***

Notar como ficou o novo banco

Salvar o novo banco

***save*** *<banconovo.dta>*

***Posicionando um banco de dados ao lado de outro***

É possível unir dois bancos acrescentando variáveis de um segundo banco, no primeiro. Por exemplo, o primeiro banco contém 3 variáveis e o segundo outras 4. É possível deixar estas variáveis em um único banco de dados. Neste caso tanto o primeiro banco como o segundo precisam conter uma variável com os mesmos números que servirá como a variável de ligação entre os bancos. Também é necessário que os bancos de dados estejam ordenados por esta variável. Esquema:

<banco1.dta>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | var1 | var2 | var3 | var4 | var5 |
| 1 | x | x | x | x | x |
| 2 | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| 100 | x | x | x | x | x |

<banco2.dta>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | var6 | var7 | var8 | var9 | var10 |
| 1 | x | x | x | x | x |
| 2 | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| . | x | x | x | x | x |
| 100 | x | x | x | x | x |

Banco resultante da combinação dos dois anteriores

<banco12.dta>

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | var1 | var2 | var3 | var4 | var5 | var6 | var7 | var8 | var9 | var10 |
| 1 | x | x | x | x | x | x | x | x | x | x |
| 2 | x | x | x | x | x | x | x | x | x | x |
| . | x | x | x | x | x | x | x | x | x | x |
| . | x | x | x | x | x | x | x | x | x | x |
| . | x | x | x | x | x | x | x | x | x | x |
| 100 | x | x | x | x | x | x | x | x | x | x |

***use*** *<banco1.dta>*

***sort*** *id*

***save,replace***

***use*** *<banco2.dta>*

***sort*** *id*

***save,replace***

***merge*** *1:1 id* ***using*** *<banco1.dta>*

Collapse

Às vezes é de interesse utilizar o banco de dados resumindo dados.

Abrir o banco <dadosrepetidos.dta>

Variáveis do banco

|  |  |
| --- | --- |
| subunidade | local de coleta |
| paisagem | tipo de paisagem |
| natureza | tipo de natureza (0 – natural, 1 – artificial) |
| ph | valores de ph das coleções de água |

Investigar como o banco está estruturado. Os níveis de ph foram aferidos várias vezes nas subunidades e é de interesse trabalhar com os valores médios de ph por subunidade.

***collapse (mean)*** *ph,* ***by****(subunidade)*

Observe como ficou o banco de dados. Utilize o Help para conhecer as demais possibilidades para resumir os dados.

Inferência estatística

***Teste de uma média com variância desconhecida***

Abrir o banco <fem.dta>

Testar a hipótese que durante a internação não ocorre mudança de peso.

***ttest*** *weight==0*

***Teste de duas médias***

Para duas amostras independentes, com pressuposições válidas (variâncias populacionais iguais e distribuição normal)

Crie uma nova variável *peso* com duas categorias (1- perdeu peso; 2 – ganhou peso)

Teste a hipótese de que a idade média das mulheres é a mesma segundo a nova variável peso.

Criando a variável peso categórica

***gen*** *peso = weight*

***recode*** *peso -4.9/0.0=1 0.1/8.3=2*

***tab*** *peso*

***label define*** *peso 1 “perdeu peso” 2 “ganhou peso”*

***label val*** *peso peso*

***ttest*** *age,****by****(peso)*

Verificando se a variável idade segue uma distribuição normal

***ladder*** *age*

***histogram*** *age,* ***by****(peso)*

Se a distribuição não for normal então é melhor utilizar o teste não paramétrico Wilcoxon (Mann-Whitney) que testa se as distribuições de idade são iguais nos grupos 1 e 2 de peso.

***ranksum*** *age,* ***by****(peso)*

***bysort*** *peso:* ***sum*** *age,* ***d***

É possível fazer o teste t de Student para uma média entrando diretamente com os dados. Considere o problema abaixo.

O nível médio de protrombina em populações normais é 20 mg/100ml de sangue. Uma amostra de 40 pacientes que tinham deficiência de vitamina K tiveram nível médio observado de protrombina de 18,5mg/100ml e desvio padrão 4mg/100ml. Seria razoável concluir que a verdadeira média de pacientes com deficiência de vitamina K é a mesma que a da população normal? Realize um teste de hipóteses.

**ttesti #obs #mean #sd #val**

**ttesti** 40 18.5 4.0 20

Conclusão:

Digitar os dados abaixo salvando o banco com o nome <salgadinhos.dta>. Defina o nome do campo que vai conter os pesos dos conteúdos dos pacotes como “peso”.

Uma companhia de produtos alimentícios utiliza uma máquina para embalar salgadinhos cujas embalagens especificam 454 gramas. Com o propósito de verificar se a máquina está trabalhando corretamente, selecionou-se 50 pacotes de salgadinhos, obtendo-se os seguintes valores de peso:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 464 | 450 | 450 | 456 | 452 | 433 | 446 | 446 | 450 | 447 |
| 442 | 438 | 452 | 447 | 460 | 450 | 453 | 456 | 446 | 433 |
| 448 | 450 | 439 | 452 | 459 | 454 | 456 | 454 | 452 | 449 |
| 463 | 449 | 447 | 466 | 446 | 447 | 450 | 449 | 457 | 464 |
| 468 | 447 | 433 | 464 | 469 | 457 | 454 | 451 | 453 | 443 |

Testar a hipótese de que a máquina está trabalhando corretamente

***ttest*** *peso==454*

Conclusão:

***Teste de duas médias com amostras independentes.***

Considerar os dados a seguir relativos ao peso ao nascer (g) de recém-nascidos com síndrome de desconforto idiopático grave. Algumas crianças foram a óbito (\*) e outras sobrevieram.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1050 | 2500 | 1890 | 1760 | 2830 |
| 1175 | 1030 | 1940 | 1930 | 1410 |
| 1230 | 1100 | 2200 | 2015 | 1715 |
| 1310 | 1185 | 2270 | 2090 | 1720 |
| 1500 | 1225 | 2440 | 2600 | 2040 |
| 1600 | 1262 | 2560 | 2700 | 2200 |
| 1720 | 1295 | 2730 | 2950 | 2400 |
| 1750 | 1300 | 1130 | 2550 | 3160 |
| 1770 | 1550 | 1575 | 2570 | 3400 |
| 2275 | 1820 | 1680 | 3005 | 3640 |

Fonte: Hand DJ *et al*., 1994.

Digitar os valores em um banco de dados com três variáveis. A primeira é “*id”*, identificação, a segunda deve conter os valores de peso ao nascer (“*pesonasc*”) e a terceira, os valores 1 para os recém-nascidos que foram a óbito e 0 para os que sobreviveram. Nomeie a terceira coluna “*status*”

Esquema:

|  |  |  |
| --- | --- | --- |
| 1 | 1050 | 1 |
| 2 | 1175 | 1 |
| . | . | . |
| . | . | . |
| . | . | . |
| 27 | 2730 | 1 |
| 28 | 1130 | 0 |
| 29 | 1575 | 0 |
| . | . |  |
| . | . |  |
| . | . |  |
| 49 | 3400 | 0 |
| 50 | 3640 | 0 |

Realize um teste de hipóteses para testar a hipótese de igualdade do peso ao nascer de recém-nascidos que sobrevivem e que vão a óbito.

***ttest*** *pesonasc,* ***by****(status)*

Conclusão:

Notar que para executar este comando, o formato do banco deve ser long.

***Teste de hipóteses de duas médias com amostras dependentes***

Se a situação envolver duas amostras dependentes, como no próximo exercício, é necessário que o banco de dados esteja em formato wide.

Esquema

|  |  |  |
| --- | --- | --- |
| 1 | 120 | 125 |
| 2 | 124 | 126 |
| . | . | . |
| . | . | . |
| . | . | . |
| 11 | 126 | 126 |
| 12 | 127 | 131 |

Primeira coluna: id (identificação do paciente)

Segunda coluna: padantes (pressão arterial diastólica antes do tratamento)

Terceira coluna: padapos (pressão arterial diastólica após tratamento)

|  |  |  |
| --- | --- | --- |
| Paciente | PAD antes | PAD após |
| 1 | 120 | 125 |
| 2 | 124 | 126 |
| 3 | 130 | 138 |
| 4 | 118 | 117 |
| 5 | 140 | 143 |
| 6 | 128 | 128 |
| 7 | 140 | 146 |
| 8 | 135 | 133 |
| 9 | 126 | 127 |
| 10 | 130 | 135 |
| 11 | 126 | 126 |
| 12 | 127 | 131 |

***ttest*** *padantes==padapos*

Conclusão:

(ANOVA)



Para fazer a análise de variância deve-se

1. Digitar o banco no formato long
2. Incluir as variáveis id, gordura, tipo

|  |  |
| --- | --- |
| gordura | tipo |
| 78 | 1 |
| 91 | 1 |
| 97 | 1 |
| 82 | 1 |
| 85 | 1 |
| 77 | 1 |
| 55 | 2 |
| 66 | 2 |
| 49 | 2 |
| 64 | 2 |
| 70 | 2 |
| 68 | 2 |
| 75 | 3 |
| 93 | 3 |
| 78 | 3 |
| 71 | 3 |
| 63 | 3 |

1. Utilizar o comando oneway gordura tipo
2. Outra forma de analisar: oneway gordura tipo, b freq means st

Conclusão:

Intervalo de Confiança

Para a média populacional, com variância desconhecida

Utilizar o banco de dados <fem.dta>

O IC(95%) para a média de QI da população de mulheres

***ci*** *iq*

É possível calcular o IC especificando os valores por exemplo, para uma amostra n=100, com média amostral =2 e desvio padrão= 0,8

***cii #obs #mean #sd***

***cii*** *100 2 0.8*

***Exercício:***

1) Construa um intervalo de 95% de confiança para estimar a pressão diastólica média populacional (), sabendo que em uma amostra de 36 adultos a pressão média amostral () foi igual a 85mmHg e o desvio padrão amostral (s) foi 12 mm Hg.

***cii*** *36 85 12*

Caso você queira outro nível de confiança, por exemplo 90%

***cii*** *36 85 12,* ***level (90)***

Conclusão

2) Uma amostra de 25 adolescentes meninos apresenta peso médio de 56 kg e desvio padrão 8 kg. Apresente o intervalo de confiança de 95% para o peso médio da população da qual esta amostra foi sorteada.

***cii*** *25 56 8*

Conclusão

Mediadas de efeito - Regressão linear múltipla

Abrir o arquivo <fem.dta>

***regress*** *weight age*

ou

Statistics> Linear regression and related > linear regression

Este comando ajusta um modelo de regressão linear de weight (variável dependente) em função de age (idade) (variável dependente, quantitativa contínua)

Desenhando a reta de regressão estimada e o IC(95%)

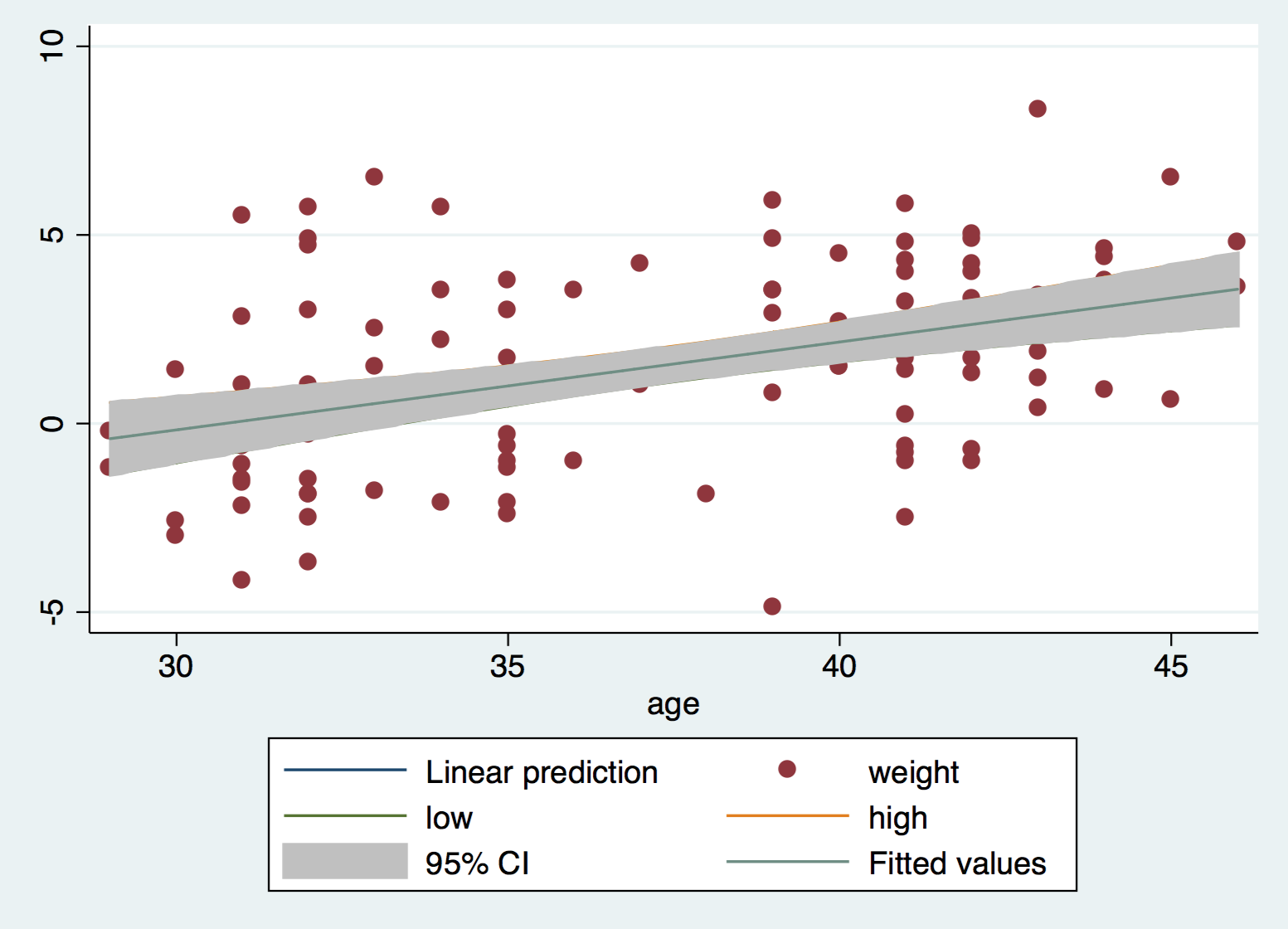
***predict*** *dp,****stdp***

***predict*** *xb,xb*

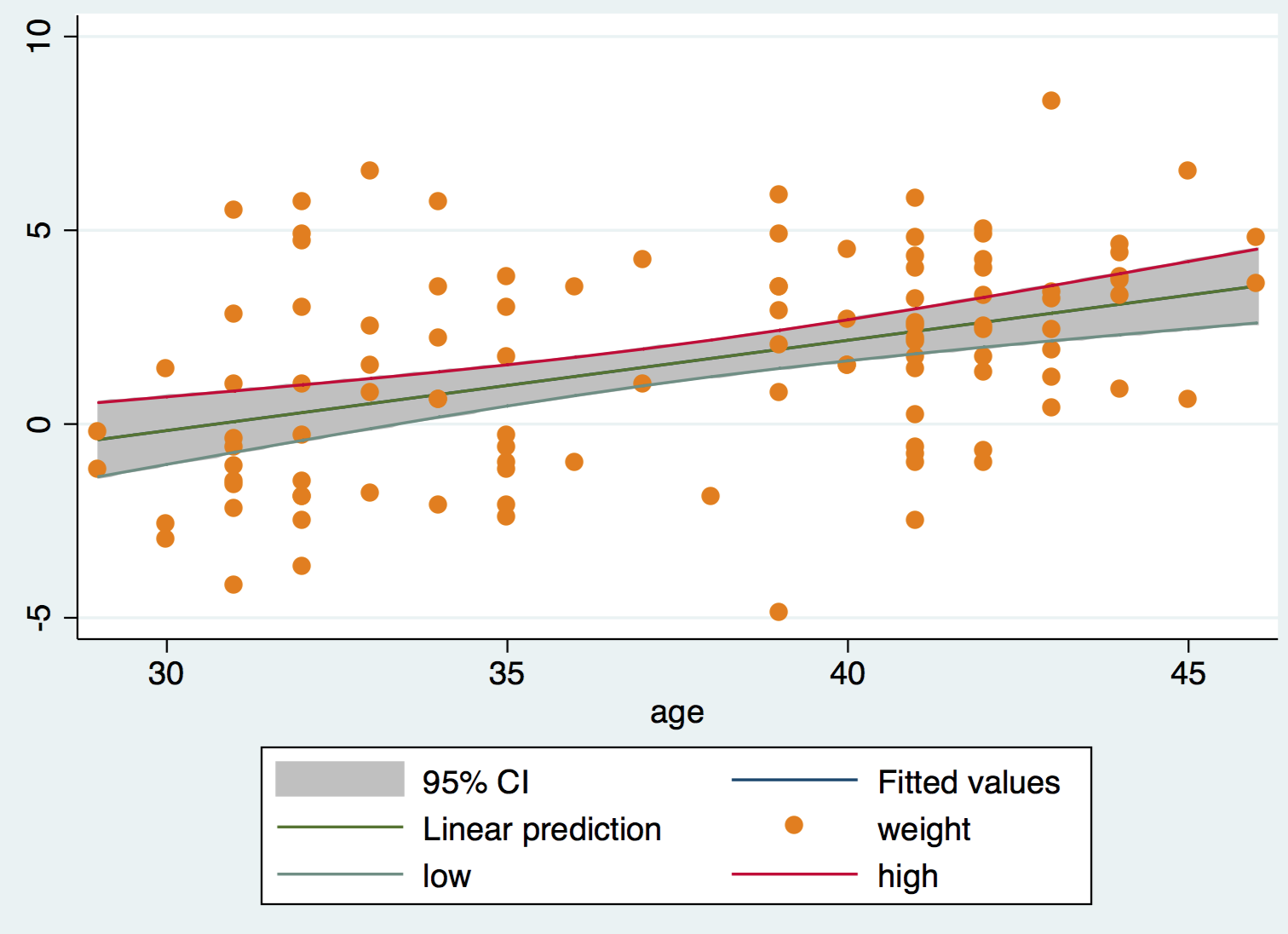
***gen*** *low=xb-1.96\*dp*

***gen*** *high=xb+1.96\*dp*

***twoway line*** *xb age|****|scatter*** *weight age||****line*** *low age, sort||* ***line*** *high age, sort||****lfitci*** *weight age*

**

***twoway lfitci*** *weight age||****line*** *xb age||****scatter*** *weight age||****line*** *low age,****sort****||****line*** *high age,* ***sort***



Incluindo a variável *depress* no modelo

***tab*** *depress,****gen****(depressd)*

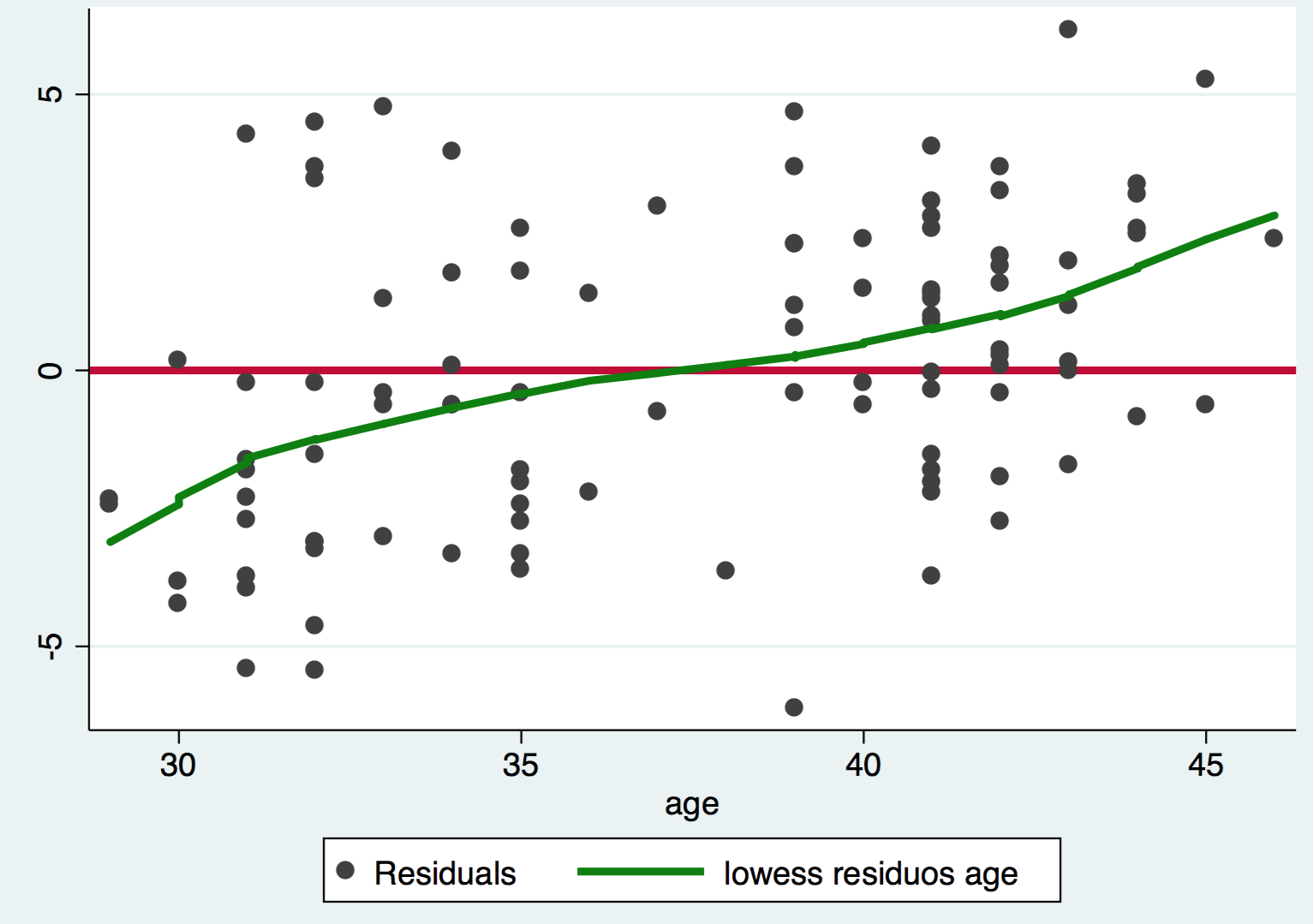
***regress*** *weight depressd2 depressd3*

***drop*** *xb*

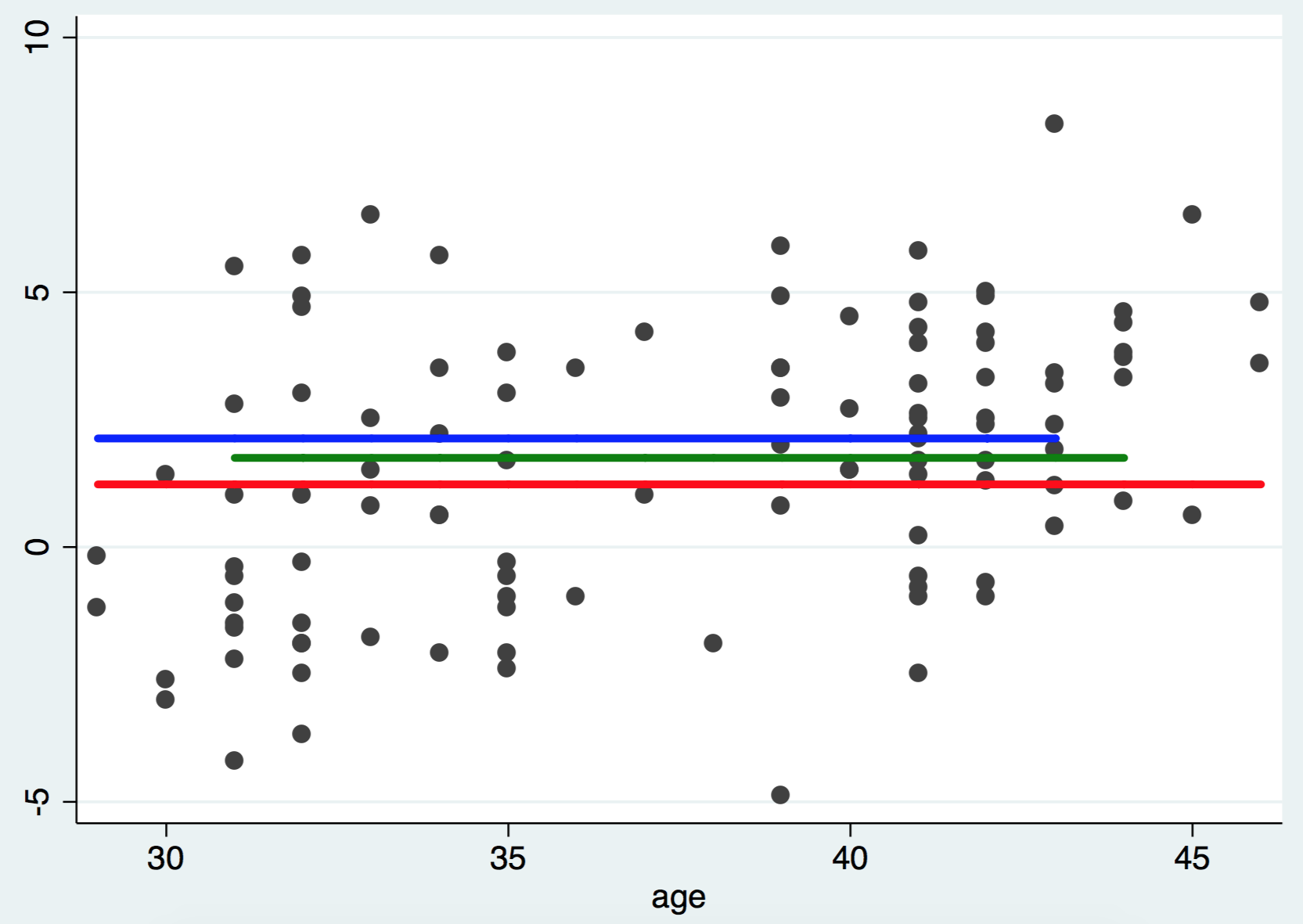
***predict*** *xb,****xb***

***predict*** *residuos,****residuals***

***scatter*** *residuos age,* ***mcolor(gs4) yline(0, lwidth(thick)****)||****lowess*** *residuos age,* ***lcolor(green) lwidth(thick)***

**

***scatter*** *weight age, mcolor(gs4) ||****line*** *xb age* ***if*** *depress==1****, lcolor(green) lwidth (thick)****||* ***line*** *xb age* ***if*** *depress==2,* ***lcolor(red) lwidth(thick)****||* ***line*** *xb age* ***if*** *depress==3****, lcolor(blue) lwidth(thick) legend(off****)*

**

Regressão logística (incidência acumulada, prevalência)

Utilizar o banco <tumor.dta>

Os dados são provenientes de um ensaio clínico aleatorizado com pacientes (0 – masculino; 1 – feminino)que tinham câncer de pulmão e foram submetidos a dois tipos de terapia de quimioterapia (0 - sequencial e 1- alternada). O desfecho (variável resposta) foi classificado em 4 categorias: 1 – piora; 2- sem mudança; 3- alguma melhora; 4 – melhora total. Dados publicados por Holtbrugge e Schumacher (1991)

***desc***

***tab*** *resposta*

***gen*** *resultado=resposta*

***recode*** *resultado 1 2=1 3 4 =0*

***tab*** *resultado resposta*

***label define*** *resultado 1“mau resultado” 0“bom resultado”*

***label val*** *resultado resultado*

***tab*** *resultado*

***tab*** *sexo*

***tab*** *sexo,****nol***

***recode*** *sexo 1=0 2=1*

***desc*** *sexo*

***label******drop*** *s*

***label define*** *sexo 1 “feminino” 0 “masculino”*

***label val*** *sexo sexo*

***tab*** *sexo*

Baixar o programa do Stata (algumas versões do Stata já terão esse pacote instalado)

**findit** *prvalue* (clicar em st0094)

Medidas de associação:

***cc*** *resultado terapia, woolf*

***cs*** *resultado terapia*

***tab*** *resultado terapia,* ***col chi***

***table*** *terapia,****c(freq mean*** *resultado)*

utilizando o comando logit que considera 1 caso (mau resultado) e 0 controle (bom resultado).

***logit*** *resultado terapia*

Comentários da saída deste comando: o coeficiente de terapia representa a diferença no log odds de resultado pior entre as terapias alternada e sequencial. O valor maior que zero indica que a terapia alternada é pior que a sequencial.

O teste de Wald tem como estatística o valor de z dado pela divisão do coeficiente pelo erro padrão.

O teste da razão de verossimilhança (LR chi2(1)) compara os modelos só com a constante e o incluindo a variável terapia (2\*(log likelihood modelo sem terapia- log likelihood modelo com terapia).

***logistic*** *resultado terapia*

***display exp****(0.4986993)*

Predizendo a probabilidade de um outcome positivo (mau resultado)

***predict*** *pr,pr*

***table*** *terapia,* ***c(freq mean*** *pr)* ***col row scol***

Análise de diagnóstico

Material em preparação

Utilizando um arquivo .do

Vamos utilizar o arquivo <sistolocinicnovo.dta>

Trata-se de amostra de 58 pacientes do sexo feminino, hipertensas, avaliadas por 6 meses.

Foram estudadas as variáveis:

|  |  |
| --- | --- |
| Variável | Labels |
| *Doença* | Doença do coração: 1 – ausente; 2- presente |
| *Droga* | tipo de medicamento utilizado no período: 1 – nenhum, 2 – tipo A; 3 – tipo B; 4 – tipo C |
| *sistolica* | incremento na pressão sistólica |
| *idade* | idade em anos |
| *salario* | renda da paciente (R$) |
| *familia* | número da família (tem pacientes da mesma família) |
| *pesoin* | peso inicial (kg) da paciente |
| *pesointer* | peso (kg) do paciente após 3 meses de tratamento |
| *pesof* | peso (kg) após 6 meses de tratamento |



Clicar no sétimo ícone (do file)

***clear***

***capture log close***

***use*** *<banco de dados>*

***log using*** *<nome do arquivo log>,* ***replace***

***desc***

***sum*** *sistólica idade salario pesoin pesointer pesof*

Criando labels

***label define*** *doença 1 “ausente” 2 “presente”*

***label val*** *doenca doenca*

***tab*** *doenca*

***label define*** *droga 1 “nenhum” 2 “A” 3 “B” 4 “C”*

***label val*** *droga droga*

***tab*** *droga*

Criando categorias para a variável idade

***gen*** *idade2=idade*

***recode*** *idade2 20/29=1 30/39=2 40/49=3 50/59=4 60/69=5 70/79=6*

***tab*** *idade2*

***label define*** *idade2 1 “20 – 30” 2 “31 – 39” 3 “40 – 49” 4 “50 – 59” 5 “60 – 69” 6 “70 – 79”*

***label val*** *idade2 idade2*

***tab*** *idade2*

Criando categorias para a variável salário

***gen*** *salario2=salario*

***recode*** *salario2 80.00/199.99=1 200.00/399.99=2 400.00/699.99=3 700.00/999.99=4 1000.00/1999.99=5 2000.00/5000.00=6*

***label define*** *salario2 1 “80.00 - 199.99” 2 “200.00 - 399.99” 3 “400.00/699.99” 4 “700.00 - 999.99” 5 “1000.00 - 1999.99” 6 “2000.00 - 5000.00”*

***label val*** *salario2 salario2*

***tab*** *salario2*

Criando ganho de peso em porcentagem

***gen*** *percentpeso=((pesof-pesoin)/pesoin)\*100*

***tab*** *percentpeso*

***sum*** *percentpeso*

***tab*** *doenca droga****, row chi***

***bysort*** *doença:****sum*** *percentpeso*

***bysort*** *droga:****sum*** *percentpeso*

Gerando variáveis dummy (indicadoras)

***tab*** *droga,* ***gen****(ddroga)*

***log close***

***clear***